BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

MPEG-4 video subjective test procedures and results

- *Pereira, F.; Alpert, T.*
Inst. Superior Tecnico, Lisbon, Portugal

This paper appears in: **Circuits and Systems for Video Technology, IEEE Transactions on**
On page(s): 32 - 51
Feb. 1997
Volume: 7 Issue: 1
ISSN: 1051-8215
References Cited: 16
CODEN: ITCTEM
INSPEC Accession Number: 5511845

Abstract:
In the recent years, the technical developments in the area of audio-visual communications, notably in video coding, encouraged the emergence of new services which are already changing our everyday life. The convergence of the telecommunications, computer, and TV/film technologies is leading to the intermixture of elements formerly characteristic of each one of these fields, creating new needs and new requirements. Among the most important trends is the need to increase the interaction capabilities between the user and the audio-visual information, notably by considering the scene as a composition of objects-the content-according to a script that describes their spatial and temporal behavior and not just a set of pixels. MPEG-4 is a new audio-visual standard aiming to establish a universal, efficient coding of different forms of audio-visual data, called audio-visual objects. To reach this target, MPEG-4 has called for proposals on techniques that may be instrumental to efficiently represent visual information, allowing simultaneously high degrees of content-based interactivity and error resilience. This paper addresses the conditions under which the proposals to the MPEG-4 first round of video subjective tests have been evaluated. Moreover, the most significative results of these tests are also presented.

Index Terms:
video coding; code standards; telecommunication standards; interactive video; audio coding; audio-visual systems; interactive video; subjective test results; audio-visual communications; video coding; audio-visual standard; audio-visual information; temporal behavior; spatial behavior; telecommunications technology; computer technology; TV/film technology; audio-visual data coding; audio-visual objects; content based interactivity; error resilience; MPEG-4 coding standard; video subjective test procedures

Reference list:

1. "Proposal package description (PPD)—revision 3", Tokyo, July 1995.

2. L. Chiariglione, "MPEG-4 project description", Jan. 1996.

3. F. Pereira, "MPEG-4: a new challenge for the representation of audio-visual information", *Picture Coding Symp. '96*, Melbourne, Australia, Mar. 1996.

4. "Systems Working Draft, version 2.0", Maceio´, Nov. 1996.

5. "MPEG-4 testing and evaluation procedures document", Tokyo, July 1995.

6. H. Peterson, "Report of the ad hoc group on MPEG-4 video testing logistics", Dallas, Nov. 1995.

7. "MPEG-4 video verification model 5.0", Nov. 1996.

8. "Recommendation ITU-R BT 812, Subjective assessment of the quality of alphanumeric and graphic pictures in teletext and similar services", 1994.

9. "Recommendation ITU-R BT 500-6, Method for the subjective assessment of the quality of television pictures", 1994.

10. "EBU Report on Recovery Time, GT V1 2651", 1994.

11. G. Bjontegaard, "H.263 anchors—technical description", Dallas, Nov. 1995.

12. "Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s.",

13. W. S. Togerson, "Theory and Methods of Scaling", *Wiley*, New York, 1958.

14. J. Ostermann, "Report on the ad hoc group on the evaluation of tools for non tested functionalities of video submissions", Dallas, Nov. 1995.

15. J. Ostermann, "Report on the ad hoc group on the evaluation of tools for non tested functionalities of video submissions for MPEG-4 in Jan. 1996", Munich, Jan. 1996.

16. W. E. Duckworth, "Me´thodes Statistiques de la Recherche Technologique", *Dunod*, 1973.

MPEG-4 Video Subjective Test Procedures and Results

Fernando Pereira, *Member, IEEE*, and Thierry Alpert

***Abstract*—In the recent years, the technical developments in the area of audio-visual communications, notably in video coding, encouraged the emergence of new services which are already changing our everyday life. The convergence of the telecommunications, computer, and TV/film technologies is leading to the intermixture of elements formerly characteristic of each one of these fields, creating new needs and new requirements. Among the most important trends is the need to increase the interaction capabilities between the user and the audio-visual information, notably by considering the scene as a composition of objects—the content—according to a script that describes their spatial and temporal behavior and not just a set of pixels.**

MPEG-4 is a new audio-visual standard aiming to establish a universal, efficient coding of different forms of audio-visual data, called audio-visual objects. To reach this target, MPEG-4 has called for proposals on techniques that may be instrumental to efficiently represent visual information, allowing simultaneously high degrees of content-based interactivity and error resilience.

This paper addresses the conditions under which the proposals to the MPEG-4 first round of video subjective tests have been evaluated. Moreover, the most significative results of these tests are also presented.

***Index Terms*—Advanced video services, MPEG-4, standardization, video subjective testing.**

I. Introduction

IN NOVEMBER 1994, at the Singapore meeting, MPEG agreed, for the first time, that "*MPEG-4 is an emerging coding standard that supports news ways (notably content-based) for communication, access, and manipulation of digital audio-visual data*" [1]. More recently, the MPEG-4 project description states that "*the MPEG-4 project aims to establish a universal, efficient coding of different forms of audio-visual data, called audio-visual objects*" [2]. After more than two years of work, starting in September 1993, MPEG-4 found its identity as the first audio-visual representation standard that understands an audio-visual scene as a composition of objects (audio, video, or audio-visual), according to a script that describes their spatial and temporal relationship. Since, in many situations, human beings complete the act of seeing by taking actions, there was the need to define a representation standard allowing the user to interact with the meaningful entities that are part of the scene (the content), overcoming the passive situation of an "only-seeing or only-hearing" user [3].

According to the MPEG-4 Proposal Package Description (PPD) [1], this "object-based approach" is also motivated by the increasing convergence between the telecommunications, computer, and TV/film technologies, leading to the mutual exchange of elements, formerly typical for each one of these areas.

To better explain its strategy, MPEG-4 defined a set of eight new or improved functionalities which are not or are not well supported by the existing or emerging standards. The eight new or improved functionalities have been clustered in three sets—content-based interactivity, compression, and universal accessibility—depending on which one they primarily address. The three sets of functionalities are not orthogonal and thus a functionality may well contain characteristics of a set in which it was not classified. The new or improved MPEG-4 functionalities are:

- Content-based interactivity
  - a) content-based multimedia data access tools;
  - b) content-based manipulation and bitstream editing;
  - c) hybrid natural and synthetic data coding;
  - d) improved temporal random access.
- Compression
  - a) improved coding efficiency;
  - b) coding of multiple concurrent data streams.
- Universal access
  - a) robustness in error-prone environments;
  - b) content-based scalability.

The current set of new or improved functionalities resulted as a compromise between the various sentiments present in MPEG at the time of its definition. These functionalities are not all equally important, neither in terms of the technical advances they promise, nor the application possibilities they open. However, they all fit the MPEG-4 vision of a framework where as many as possible types of audio-visual data [natural, synthetic, mono, stereo, multichannel, two-dimensional (2-D), and three-dimensional (3-D)] are represented, accessed, and manipulated.

Moreover, the rapidly evolving technological environment of the last years showed that standards which set a specific algorithm and the corresponding syntax and do not take into account the continuous development of the hardware and of the methodologies, risk becoming obsolete relatively soon. Thus, flexibility and extensibility are essential features for a

Manuscript received March 10, 1996; revised July 1, 1996. This paper was recommended by Guest Editors Y.-Q. Zhang, F. Pereira, T. Sikora, and C. Reader. This work was supported by CEC under the RACE program—project MAVT R2072 and the ACTS program—project TAPESTRIES AC055. The work of F. Pereira was supported by the Junta Nacional de Investigação Científica e Tecnológica under the project "Processamento Digital de Audio e Vídeo."

F. Pereira is with the Instituto Superior Técnico/Instituto de Telecomunicações, 1096 Lisboa Codex, Portugal.

T. Alpert is with the Centre Commun d'Etudes de Télédiffusion et Télécommunications, Cesson Sévigné Cedex, France.

Publisher Item Identifier S 1051-8215(97)00881-1.

standard in the current moving technological landscape and should be provided in MPEG-4 by the MPEG-4 Systems and Description Languages (MSDL) [4]. The MSDL plays the role of MPEG-4 system layer defining a complete framework for supporting flexibility, extensibility, and the various MPEG-4 functionalities of content-based manipulation, efficient compression, universal access, and other standard functionalities. The MSDL will allow the conveyance to a decoder of a particular assembly of the MPEG-4 tools to satisfy specific user requirements as well as to describe new algorithms and download their description to the decoding processor for execution [2].

When the functionality-based approach was adopted in MPEG-4, it became clear that testing would be a difficult task for testing experts, since there was no significant experience for the tests and conditions that were foreseen—object quality, error resilience, very low bit rates. Since the MPEG work usually proceeds by issuing calls to gather technology that will be tested, the work on the definition of the appropriate procedures to evaluate the proposals that would arrive at MPEG started immediately.

One of the new challenges in MPEG-4 is the request to simultaneously address more than one target/functionality depending on the set of requirements being considered. In fact, it is expected that MPEG-4 will allow the simultaneously provision of a few functionalities which will very likely place opposing requirements. This imposes the need not only to optimize the use of the set of tools providing each functionality but also to find the good compromise between the algorithms addressing a cluster of functionalities which should be instrumental for as many applications as possible.

According to the current MPEG-4 workplan, two rounds of tests are foreseen. The first round of tests was performed in two dates, November 1995 and January 1996. However, formal video subjective tests were only performed between 30 October and 3 November 1995, in Los Angeles, at the premises of the Hughes Aircraft Company. The second round of MPEG-4 tests is foreseen for July 1997.

This paper intends to address the video subjective tests procedures under which the November 1995 video subjective tests were conducted. Moreover, the most relevant results of these tests will be presented. These two topics are detailed in [5] and [6]. Since this special issue addresses the MPEG-4 video proposals, the inclusion of a paper with the description of the corresponding test procedures, and thus of the constraints under which the proposals were submitted, is essential to fully understand their technical content.

Regarding the assessment environment, only the conditions indicated in advance to the tests in the "MPEG-4 Testing and Evaluation Procedures Document" will be shortly mentioned. Due to logistic or other relevant reasons, some small changes were made in the testing specification defined in the "MPEG-4 Testing and Evaluation Procedures Document." These changes were agreed in the context of the Ad Hoc Group on MPEG-4 Testing Logistics, set at the MPEG Tokyo Meeting, held in July 1995, which had the mandate to address the last minute details related to the testing logistics.

A. Testing and Evaluation of Tools and Algorithms

In the context of MPEG-4, a tool is a technique that is accessible via the MSDL or described using the MSDL [1]. An algorithm is an organized collection of tools that provides one or more functionalities [1]. Tools and algorithms were the technical elements that MPEG asked for and tested or evaluated in order to reach the identified targets.

Taking into account the type of functionalities and the limitations at the time of the tests (e.g., no hardware was available, which prevents testing interactive use), two types of tests were envisaged for the video proposals:

- Subjective testing—conventional subjective viewing tests, including bit rates lower than usual;
- Evaluation by experts—evaluation, by a group of experts, of the (description of the) coder's implementation, and assessment of how well the requested functionalities are supported; it was expected to be used when limitations prevented more direct testing.

According to the "MPEG-4 Testing and Evaluation Procedures Document," relevant criteria for the overall evaluation were: formal subjective tests results, efficacy in addressing all the MPEG-4 new, improved, or standard functionalities [1], added value, for example in addressing functionalities not referred in the MPEG-4 PPD, adaptability to relevant changing conditions such as different input (scene content), different (number of) objects, different and/or variable bit rates, different and/or changing error patterns, different temporal and spatial resolutions or other kinds of adaptability, margins for improvement, and complexity.

The testing and evaluation procedures for tools and algorithms specified for the first MPEG-4 round of tests were strongly conditioned by the intrinsic difficulties associated with the testing of tools and the need to keep the overall tests manageable. This led to the need to choose a representative set of the new or improved MPEG-4 functionalities to be more fully tested and evaluated. However, although only part of the possible tests was performed in the first round of tests, proposals for all the functionalities were welcome.

The new or improved MPEG-4 functionalities chosen as representative functionalities for full testing were: content-based scalability, improved compression, and robustness in error-prone environments. Content-based scalability refers to the ability to achieve a scalable representation (base + enhancements layers) of video information with fine granularity in content, spatial resolution, temporal resolution, or SNR, on an object basis. Improved compression refers to the target of providing subjectively better audio-visual quality compared to existing or other emerging coding standards (such as ITU-T H.263), at comparable bit rates. Finally, robustness in error-prone environments refers to the provision of error resilient video streams in order that they can be used over a large variety of networks with possibly severe error conditions.

Due to the difficulty of formally testing and evaluating tools, it was decided they would be evaluated by a panel of experts. The algorithms addressing the three representative functionalities were tested by means of formal subjective testing and, sometimes, also by a panel of experts. All algorithms

TABLE I
MPEG-4 VIDEO LIBRARY

Sequence class	Input library	Content complexity	Video test material
A	ITU-R 601	Low spatial detail and low amount of movement	Mother & Daughter (60 Hz), Akiyo (60 Hz), Hall Monitor (60 Hz), Container Ship (60 Hz), Sean (60 Hz)
B	ITU-R 601	Medium spatial detail and low amount of movement or vice versa	Foreman (50 Hz), News (60 Hz), Silent Voice (50 Hz), Coast Guard (60 Hz)
C	ITU-R 601	High spatial detail and medium amount of movement or vice versa	Table Tennis (60 Hz), Stefan (60 Hz), Mobile & Calendar (60 Hz), Fun Fair left view (50 Hz)
D	2 * ITU-R 601	Stereoscopic	Tunnel (50 Hz), Fun Fair (50 Hz)
E	ITU-R 601	Hybrid natural and synthetic	Children (60 Hz), Bream (60 Hz), Weather (60 Hz), Destruction (60 Hz)

addressing nonrepresentative functionalities were dealt with in the same way as tools; this means they were evaluated by a panel of experts.

The proposers of algorithms addressing the three representative functionalities were basically required to provide a technical description (with bitstream statistics), a decoder executable, and the bitstreams and decoded sequences (in D1 tape) corresponding to the test cases addressed by the proposal.

In the following, we will mainly address the test procedures and conditions regarding the formal video subjective testing of the three representative functionalities. Section II will describe the test material and the conditions under which the tests were performed. Section III will introduce the subjective tests used. In Section IV, the testing procedures for the three representative functionalities will be presented. Section V will describe how the results were processed, and Section VI will present the main results of the tests. Finally, Section VII will comment on the main lessons learned with these tests.

II. TEST MATERIAL AND CONDITIONS

In the following sections, the common characteristics for the test material and for the testing conditions related to the MPEG-4 first round of tests will be presented.

A. Source Material and Bit Rates

For the MPEG-4 video tests, an input library was established containing all the source data. The input library contained video sequences in either 60 Hz ITU-R 601 format or 50 Hz ITU-R 601 format. These formats constituted the highest quality of data used in the tests.

The sequences in the input video library were classified into five classes according to their characteristics in terms of spatial detail and movement, as shown in Table I. This table also indicates the specific sequences that were chosen for each class.

For all the video sequences indicated in Table I, the first 300 frames in the input library were to be considered the input material. Exceptions were: "Sean" where frames 2 to 301 (starting with 1) were to be considered and "Fun fair left view" where the 300 frames were obtained by palindroming the last 50 frames of the 250 frames distributed (this means a sequence of 1, ···, 250, 249, ···, 200 with fields reordered where needed).

Since most of the new or improved MPEG-4 functionalities are content-based, the segmentation of the video data into "objects" was needed. To the maximum extent possible (and this happened, at least, for all the sequences used in the content-based interactivity tests), the video input library was made available with segmentation information that proposers could choose to use or not. Any other segmentation method was allowed, provided it was included in the proposal technical description.

The sequences were made available according to the formats described in the following.

1) Class A, B, and C Sequences: For class A, B and C sequences, segmentation could have a maximum of 256 segments. Whenever possible, the segments should have a semantic meaning. For these classes of sequences, two types of files were used (class D sequences have two views/sets of files similar to class A, B, or C sequences):

1) Luminance and chrominance (YUV)—ITU-R 601 format containing luminance and chrominance data (all frames were chained without gaps, with Y, U, and V data chained for each frame).
2) Segmentation—The format for the exchange of the segmentation information was similar to the one used for the

TABLE II
VIEWING CONDITIONS

Viewing distance	4H (*)
Peak luminance	from 70 to 250 cd/m^2 (indirectly deduced by the screen contrast ratio)
Screen contrast ratio	from 30 to 50
Ratio of background luminance to maximum screen luminance	~ 0.25
Illumination in the viewing area	200 lux
Screen size	around 17" in diagonal

(*) H indicates the screen height; therefore the distance between the subject and the screen had to be four times the screen height, i.e., approximately 1 m for a 17" screen.

images, i.e., a segmentation mask had a format similar to ITU-R 601 luminance, where each pixel has a label identifying the segment it belongs.

2) Class E Sequences: Class E sequences are composed of natural and synthetic content and they are understood as a composition of several layers (segments/objects).[1] Thus, class E sequences data was provided through two types of files.

1) Luminance and chrominance (YUV) files—ITU-R 601 format containing the luminance and chrominance data in the same format as for the class A, B, and C sequences. One file per layer.
2) Alpha planes (alp) files—One ITU-R 601 format file per layer containing the corresponding alpha values. The alpha-planes represent the blending contribution of each layer for every part of the scene. For the layered composition of a sequence, each layer has its own YUV and alpha files.

The alpha values, using a range of [0, 255], were provided in one of two possible formats—format A using premultiplied alpha and format B using nonpremultiplied alpha—described in detail in [5].

Since the input library was available in either 50 Hz or 60 Hz ITU-R 601 formats and some of the bit rates to be tested were quite low, the need for downsampling and upsampling was foreseen. Thus, some downsampling and upsampling filters are suggested in [5]. Alternative downsampling and upsampling filters could also be used provided a precise report was included in the proposal technical description.

The following bit rates were considered basic for the MPEG-4 video testing: 10, 24, 48, 112, 320, 512, and 1024 kb/s. The bit rate was computed as the total number of bits divided by the length of original sequence in time (including the first frame).

B. Presentation and Viewing Conditions

The sequences were assessed after prerecording onto video tape. The source recordings were processed off-line through the systems under test and the output was rerecorded to provide the set of recordings to be evaluated. For the editing of the test tapes, a digital VTR (D1) was used to minimize the impairments that were not produced by the system under test.

The display format was always the 60 Hz ITU-R 601 format. This format was either full or windowed, depending on the particular conditions of the tests (see Section IV). The details on the display format for the various classes of video sequences are described in [5].

The sequences provided for the tests had to be already in the display format indicated in [5], following the conditions indicated, for each test, in Section IV. The 50 Hz sequences were looked at as 60 Hz sequences and thus all computations, notably those related to the bit rate, were done in 60 Hz conditions.

The set of viewing conditions in Table II was suggested by Rec. ITU-R 812 [8] for the setup of the laboratories [5]. However, at the tests in Los Angeles, basically the Rec. ITU-R 500 was used.

III. DESCRIPTION OF SUBJECTIVE TEST METHODS

The subjective test methods used during the MPEG-4 video subjective tests are briefly described in this section. Depending on the characteristics of the functionality being tested, different test methods have been chosen: single stimulus, double stimulus impairment scale, double stimulus continuous quality scale, or double stimulus binary vote methods.

The subjects participating in the formal subjective tests were chosen among those with normal color vision and normal or corrected-to-normal visual acuity. Only MPEG experts were used in the tests, and each group of assessors constituted 15 experts.

The sequences assessed were 10 s long. The time for voting was set to 10 s, since voting sheets were used and no automatic systems for voting collection were available.

Beside the effective testing phase, two other phases could be distinguished in the time alignment of the sessions: the training phase before the assessment phase started and the stabilization phase already after the testing phase started (Fig. 1).

For each testing session, the training phase should last about 5 min. The training phase has the main objective of giving instructions to the assessors, usually in written form. At the

[1] According to the current MPEG-4 terminology, a layer is now called a video object plane (VOP) [7].

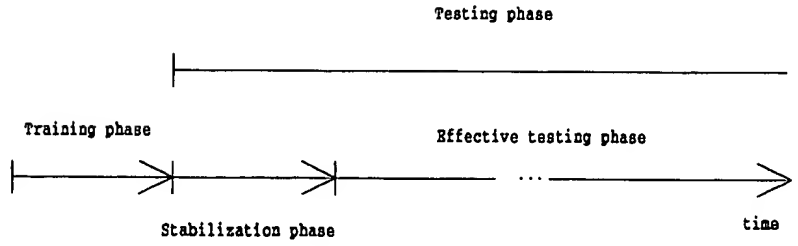


Fig. 1. Phases of a subjective testing session.

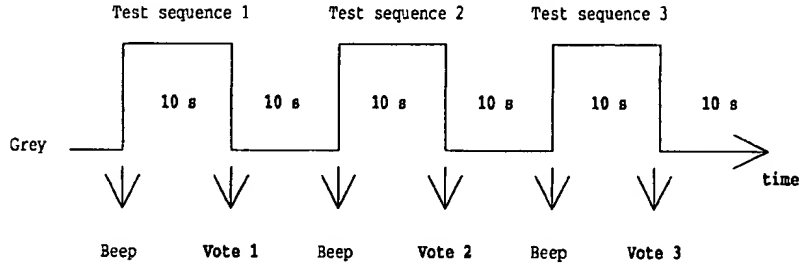


Fig. 2. Presentation sequence for the SS method.

MPEG-4 tests, the assessors were given written instructions which, after some time for reading, were also briefly explained, mentioning the type of assessment, the opinion scale, and the presentation of the stimuli. The assessors were also given the opportunity to ask questions. Answering questions about the procedures or about the meaning of the instructions was done with care to avoid bias.

During the training phase, a trial with two or three presentations was done, to get the assessors used to the timing and the quality range shown during the test. When the same assessment method was used in subsequent test sessions with the same group of assessors, the training phase was made only before the first test session.

The main goal of the stabilization phase is to "stabilize" the assessors' judgements by making them vote quality cases corresponding to the range of quality shown during the test. This phase consists of the first five presentations of each test session, including some of the best and worst conditions, so that the whole impairment range is shown. The votes corresponding to these presentations were not taken into account and the same presentations were proposed again during the test session. The assessors should not be aware of this phase. Each testing session should last no more than about 30 min and the same group of assessors should be involved in the tests for no more than 2 h per day.

To avoid confusion between sequences or between sequences and references, it was considered convenient for the observers to have some audio cues which will be indicated in the following for each test method.

A. Single Stimulus Method (SS)

This method is based on the single stimulus method described in Rec. ITU-R 500 [9]. Fig. 2 shows the typical presentation sequence for the SS method, including the audio cues below the arrows. For the MPEG-4 tests, a few modifications were suggested in order to tailor the method to take into account the constraints related to lower bit rates. The main topics that had to be specified were the grading scale and the anchoring.

The results of quality assessments often depend not only on the actual video quality, but also on other factors such as the total quality range of the test conditions, the experience and expectations of the assessors, etc. In order to control some of these effects, a number of dummy test conditions were added and used as anchors. These anchors are usually well-known conditions, for example, available coding standards (see Section IV).

Rec. ITU-R 500 recommends five-level categorical scales for either quality or impairment evaluation. However, recent studies show that numerical scales can provide better results mainly because they can remove the bias due to the interpretation of the categories that describe the quality levels. In particular, the 0-10 numerical scale seems to increase the stability of the results [10]. Thus an 11-level numerical scale was used for the SS tests performed in the context of the MPEG-4 first round of testing.

B. Double Stimulus Impairment Scale Method (DSIS)

This method is an adaptation of the ITU-R Double Stimulus Impairment Scale method taking into account the use of lower bit rates. This use is mainly reflected in the presentation and reference conditions.

The subjects were first presented with a reference sequence, then with the processed version of the sequence. Following this, they were asked to vote on the impairment of the second sequence with respect to the first one, using a five-level im-

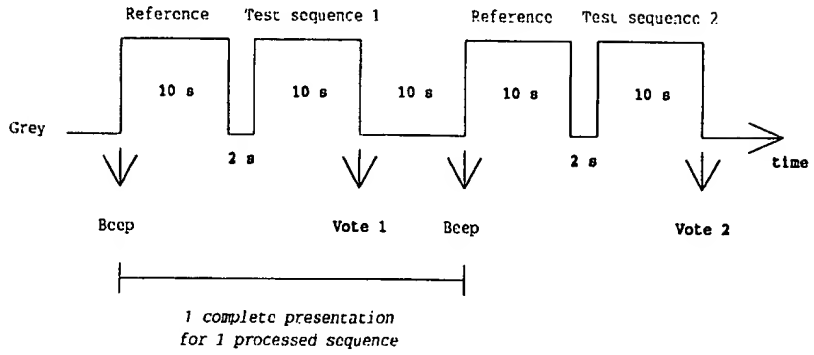


Fig. 3. Presentation sequence for the DSIS and DSBV methods.

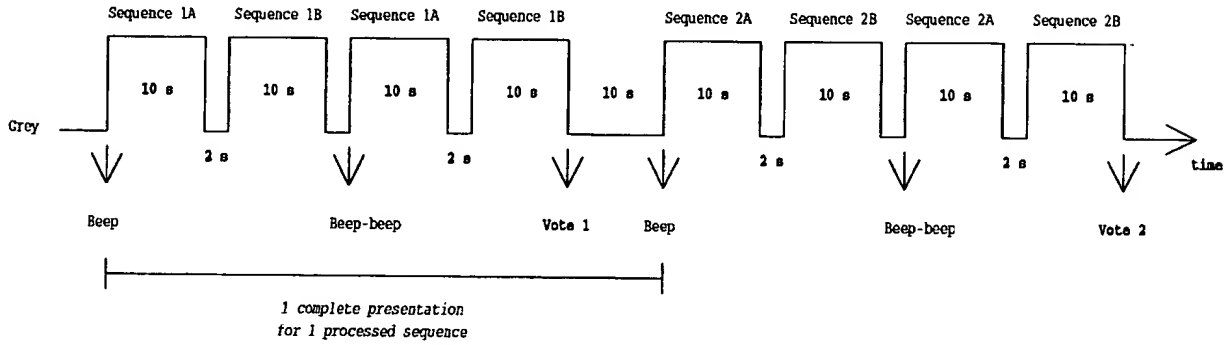


Fig. 4. Presentation sequence for the DSCQS method.

pairment scale—*very annoying, annoying, slightly annoying, perceptible but not annoying, imperceptible.*

In order to reduce the duration of the test, each pair of sequences was presented just once. The time for displaying of the conditions and for voting was 10 s, while the interval between the reference and the corresponding test condition was 2 s (see Fig. 3).

The pairs of references and coded sequences were presented in a pseudorandom order. In any case, the same sequence was never presented on two successive occasions with the same or different levels of processing. For the MPEG-4 tests, the original sequences, at the convenient spatial resolution, were used as references.

C. Double Stimulus Continuous Quality Scale Method (DSCQS)

The Double Stimulus method with a continuous quality scale was taken from Rec. ITU-R 500 [9]. All the sequences are presented unimpaired (assessment reference) and impaired. The basic principle is to assess pairs of sequences. One of the two sequences is a reference according to the basic formats previously identified; the other has been subject to processing by the system being assessed. Observers are not informed of the relative position of these two sequences and they have to allocate a score to each individual sequence using a continuous quality scale. Each couple of sequences is repeated once. The time sequencing of this method is illustrated in Fig. 4.

The DSCQS method has been defined in order to provide a classification for the various systems tested. It therefore provides relative quality assessments. One means of approaching more absolute results is to introduce systems based on known quality into the test. These high or low quality systems are the anchors already referred for single stimulus tests. In addition to a ranking, the results then show a measure compared to the anchor. The DSCQS method gives results that are context-related, by its very principle, and thus no absolute assessment compared to an ideal image can be attained. However, it seems to be relatively stable for classification purposes.

While the DSCQS method is normally used when the references and the sequences to test have similar qualities (their relative presentation order is not known), the DSIS method is normally used when the reference has clearly a higher quality and thus the sequences to test are evaluated with respect to the reference (always displayed in first place).

D. Double Stimulus Binary Vote Method (DSBV)

This method is similar to the DSIS method in terms of presentation sequence. The first sequence displayed is a reference sequence, and the second sequence displayed is a sequence to which was applied a negative effect that should have as small as possible an impact on the quality of the sequence at

TABLE III
SPECIAL OBJECTS FOR CONTENT-BASED SCALABILITY TESTS

Class of sequence	Sequence	Special objects for scalability tests
A	Akiyo	Woman
	Scan	Man
	Hall Monitor	Man with monitor
B/C	News	Dancers in the monitor
	Coast Guard	Big boat
	Stefan	Tennis player
E	Weather	Woman
	Children	Kids
	Bream	Fish
B/C/E	Stefan	Tennis player
	Children	Kids
	News	Dancers in the monitor
	Coast Guard	Big boat

the time of the voting. The voting is binary in the sense that the assessor has just to vote if the second sequence recovered or not from the negative effect applied (in comparison with the first one to which no negative effect was applied), by the moment of the voting.

In the MPEG-4 tests, the negative effect was the corruption of the coded bitstream by some errors following a previously identified distribution. The idea was thus to compare, by the end of the sequence (recovering time), the reference sequence (decoded image without application of errors) and the sequence decoded from the corrupted bitstream. A "YES" answer should be given to the cases where the corrupted sequence recovers from the errors and a "NO" answer should be given to the cases where recovery was not evident.

IV. VIDEO SUBJECTIVE TESTING PROCEDURES

This section will address the specification of the MPEG-4 video subjective tests for the algorithms addressing the three representative functionalities. In fact, due to the need to keep the amount of tests manageable, only some representatives of the MPEG-4 new or improved functionalities were fully tested. This basically means that not all the functionalities were tested in an independent way. The three sets of tests performed reflect the need to consider all the main functionality areas of MPEG-4 together with the absolute need to have a credible but simultaneously realistic set of tests performed.

A. Content-Based Scalability

Content-based scalability is the new MPEG-4 functionality chosen to represent the content-based interactivity capabilities in the first round of tests. According to the MPEG-4 PPD, "*MPEG-4 shall provide ability to achieve scalability with a fine granularity in content, quality (spatial resolution, temporal resolution) and complexity*" [1].

The goal of content-based scalability is to achieve bitstream and complexity scalability. Bitstream and complexity scalability can be achieved by scaling the content (object) and the quality of the audio-visual information (on an object basis). Quality scalability can be achieved by scaling the spatial or temporal resolutions (spatial and temporal scalabilities).

1) Test Material and Conditions: Since it would be desirable that the results of the content-based scalability tests not depend on the segmentation processing, MPEG made available source material that was already segmented into distinguishable objects. However, the use of other segmentation masks was allowed provided their description was made available in the proposal technical description.

For these scalability tests, four classes of test material, in ITU-R 601 format, were used according to the following classification:

- Class A—restricted sequences with low activity and small number of objects;
- Class B/C—generic scenes with higher activity and more objects;
- Class E—sequences containing hybrid synthetic and natural content;
- Class B/C/E—sequences either only with natural content or combining natural and synthetic content.

The sequences used for the content-based scalability tests are indicated in Table III.

TABLE IV
OBJECT SCALABILITY TEST PARAMETERS AND CONDITIONS

Object Scalability							
		Format					
Sequence class	Bitrate kbps	Library(1)	Reference	Display	Test material (4)	Test method	Anchor
A	48	ITU-R 601	no reference	CIF (2)	3 sequences of 10 s	SS	H.263 at same bitrate(2)
B/C	1024	ITU-R 601	ITU-R 601	60 Hz ITU-R 601	3 sequences of 10 s	DSCQS	MPEG-1 at same bitrate(3)
E	320	ITU-R 601	ITU-R 601	60 Hz ITU-R 601	3 sequences of 10 s	DSIS	MPEG-1 at same bitrate(3)

(1) Presegmented sequences; (2) CIF displayed in a window centered in the 60 Hz ITU-R 601 format; (3) CIF upsampled to 60 Hz ITU-R format; (4) See Table III.

Table III also indicates the objects in the sequences for which some type of scalability will have to be applied (for example, in the sequence Children, the kids will appear with higher spatial or temporal resolution, if spatial or temporal scalability is used).

2) Test Methods: In the first set of MPEG-4 tests, object and quality scalability were tested subjectively. Object scalability was also evaluated by a panel of experts.

Proposals had to address only one type of scalability at a time—object or quality scalability. For quality scalability, only one type of resolution—spatial or temporal—could vary between the two scaleable layers provided—base and enhancement layers. Proposers had complete freedom about the coding resolutions to use as well as about the pre- and postprocessing techniques. The only format restrictions were associated with the input library and display formats indicated in Tables IV and V. However, the proposal technical description was required to clearly indicate all the details about the coding resolutions and the pre- and postprocessing techniques used.

Even though MPEG-4 wants this functionality to support scalability at a fine granularity, only two layers of quality scalability had to be provided in the bitstreams to be tested in the first round of tests. However, solutions which support fine granularity scalability were encouraged.

A proposal had to address, at least, a row in Table IV or a row in Table V. This means that a proposal had, at least, to provide bitstreams and coded sequences for all the sequences of a certain sequence class. The content-based scalability tests performed in the first MPEG-4 set of tests are described in the following.

a) Object scalability: Object scalability was subjectively tested and evaluated by a panel of experts. Object scalability is associated with the capability to control the number of simultaneous objects decoded and displayed.

The test sequences (all objects) had to be coded at the bit rate specified in Table IV for each class of sequences without any constraints on the coding resolutions. The tests evaluated the subjective quality of the total image. The ability to decode a subset of the objects using a subset of the bitstream as well as the ability to build the image object by object were evaluated by a panel of experts. Table IV describes the test parameters and conditions for object scalability tests; the distribution of bit rate among the various objects in a sequence was free. The sequences used for each class as well as the selected objects for each sequence, are specified in Table III.

In order to help the experts' evaluation, proposers were required to provide a D1 tape with a demo, showing the performance of the algorithm proposed in terms of object scalability. For each sequence, this tape had to show the selected object indicated in Table III, displayed alone, coded in the same conditions used for the testing of the total image quality (this means the same division of the bit rate among the various objects in the sequence). The proposal technical description was asked to indicate the number of bits used to code the selected object. The object had to be shown alone, during 10 s, in a grey background (Y = Cr = Cb = 128).

The precise conditions in which the ITU-R H.263 anchors were generated are described in [11]. Basically, all the advanced coding options were used for all the sequences. PB-frames were switched on and off according to an adaptive scheme described in [11].

MPEG-1 anchors were generated in a compatible way with the specifications in ISO/CEI DIS 2-11 172—MPEG-1 [12].

b) Quality scalability: Quality scalability may be achieved through spatial or temporal resolution scaling. Since

TABLE V
SPATIAL AND TEMPORAL SCALABILITY TEST PARAMETERS AND CONDITIONS

Spatial and Temporal Scalability						
		Format				
Sequence class	Bitrate kbps	Library[1]	Reference	Display	Test material [3]	Test method
A	24 (base)	ITU-R 601	no ref.	CIF[2]	3 sequences of 10 s	SS
	24 + 24 (base + enhancement)	ITU-R 601	no ref.	CIF[2]	3 sequences of 10 s	SS
B/C/E	512 (base)	ITU-R 601	ITU-R 601	60 Hz ITU-R 601	4 sequences of 10 s	DSCQS
	512 + 512 (base + enhancement)	ITU-R 601	ITU-R 601	60 Hz ITU-R 601	4 sequences of 10 s	DSCQS

(1) Presegmented sequences; (2) CIF displayed in a window centered in the 60 Hz ITU-R 601 format; (3) See Table III.

only two layers were used in the tests (a base layer plus only one enhancement layer), the proposers were requested to scale the information by using only one of the mechanisms referred to below. Proposers were requested to clearly describe the mechanism they used to scale the information for the special objects previously identified.

The test sequences (all objects) had to be coded at the bit rates specified in Table V, for each class of sequences, without any constraints on the coding resolutions. The quality of the selected object(s) was changed independently of the rest of the scene using the additional bit rate indicated in Table V (enhancement layer). The selected objects whose quality had to be improved are indicated in Table III.

Different approaches were adopted to assess the quality of the two layers, both for spatial and temporal scalability. The global quality of the base layer was tested through a formal subjective test performed on the sequences coded with the base bit rate. The enhancement layer was tested through a formal subjective test of the quality of the selected object(s) whose spatial or temporal resolution had been scaled by using the additional bit rate. For the formal subjective testing of the enhancement layer, observers had to be informed about the selected object(s) for each sequence in which they had to concentrate their attention.

Table V describes the test parameters and conditions for content-based quality scalability testing.

Spatial Scalability: The spatial resolution of the selected object(s) was changed independently of the rest of the scene. A base bit rate was used to encode the entire sequence. An additional bit rate in the enhancement layer was used to enhance the designated object(s). This means that spatial scalability is applied when, after choosing a certain combination of spatial/temporal resolutions for each object in the base layer (maybe all objects are coded with different resolutions), the spatial resolution of some selected objects (indicated in Table III) is improved using the bit rate associated to the enhancement layer. A formal subjective test was performed on the global quality of the sequence coded with the base bit rate and another formal subjective test was performed on the quality of the selected object(s) whose spatial resolution had been scaled in the sequence coded with the additional bit rate.

The spatial resolution is here generically associated with the representation of an object at a certain moment in time; this means that spatial resolution may be associated with the number of pixels in the object, with the accuracy of these pixels, or other parameters, such as the discrete cosine transform (DCT) coefficients.

Temporal Scalability: The temporal resolution of the selected object(s) was changed independently of the rest of the scene. A base bit rate was used to encode the entire sequence. An additional bit rate in the enhancement layer was used to enhance the designated object(s). This means that temporal scalability is applied when, after choosing a certain combination of spatial/temporal resolutions for each object in the base layer (maybe all objects are coded with different resolutions), the temporal resolution of some selected objects (indicated in Table III) is improved using the bit rate associated with the enhancement layer. A formal subjective test was performed on the global quality of the sequence coded with the base bit rate and another formal subjective

TABLE VI
TEST MATERIAL AND CONDITIONS FOR COMPRESSION TESTING

Sequence class	Number of sequences	Test material	Bitrate (kbps)
A	4	Mother & Daughter; Akiyo; Hall Monitor; Container Ship	10, 24, 48
B	4	Silent Voice; Foreman; News; Coast Guard	24, 48, 112
C	4	Fun Fair; Table Tennis; Mobile & Calendar; Stefan	320, 512, 1024
E	4	Weather; Children; Bream; Destruction	48, 112, 320

test was performed on the quality of the selected object(s) whose temporal resolution had been scaled in the sequence coded with the additional bit rate.

B. Compression

Improved coding efficiency was the functionality chosen to represent the MPEG-4 compression capabilities in the first round of tests. According to the MPEG-4 PPD, the coding algorithms that will be developed in the framework of MPEG-4 are expected to provide "*subjectively better audio-visual quality at comparable bit rates compared to existing or other emerging standards*" [1]. Thus, for this set of tests, the formal subjective tests should be carried out not only to rank the candidates, but also to compare their compression efficiency with that of other available or emerging standards.

The formal subjective tests had to take into account a wide range of bit rates, that means a wide range of subjective quality, picture complexity, etc. The assessment strategy described in the following section was intended to fulfil the above two points and, at the same time, to keep the number of tests manageable.

1) Test Material and Conditions: For the compression tests, sequences with different content complexity were used. The video sequences used in the tests and the corresponding conditions are indicated in Table VI.

Each proposal may address one or more classes of sequences. However, for each class addressed, all bit rates and all sequences for that class (as shown in Table VI) must be processed.

2) Test Methods: Taking into account the foreseeable characteristics of the video signal coded at the various bit rates, different test methods were used for different ranges of bit rates. In order to compare the performance of the new coding schemes with respect to the existing or emerging standards, suitable anchor conditions were used for each range of bit rates. These anchor conditions were produced by coding the test sequences with standard schemes at the same bit rates used to produce the test material. For each range of bit rates, the standard expected to perform the best was used.

Table VII summarizes the test methods and test conditions used for each range of bit rates. Proposers had complete freedom about the coding resolutions to use as well as about the pre- and postprocessing techniques (provided they were described in the proposal technical description).

To keep the first set of tests manageable, proposals addressing class D sequences were evaluated in the same way as tools (evaluation by a panel of experts). Formal subjective testing is foreseen at a later stage.

C. Robustness in Error Prone Environments

Robustness in error-prone environments was the MPEG-4 functionality chosen to represent universal access capabilities in the first round of tests. According to the MPEG-4 PPD, "*MPEG-4 shall provide an error robustness capability to allow access to applications over a variety of wireless and wired networks and storage media*" [1].

It was the purpose of this test to determine how well a codec operates in error-prone environments by subjecting the compressed bitstreams to residual error conditions[2] representative of those conditions provided by a variety of networks. To allow for operation in the presence of these residual error conditions, MPEG-4 codecs should incorporate error control methods, such as forward error correction, error containment, and error concealment. Two categories of tests were used: error resilience and error recovery.

1) Test Material and Conditions: Many of the methods used to provide error robustness to a codec introduce additional overhead in the compressed bitstream. In order to allow the greatest flexibility to the codec developers, there were no restrictions to the use of forward error correction, error detection, resynchronization, etc. within the compressed bitstream. For the error-resilience and error-recovery proposals, for each decoded sequence, only the

[2]The residual errors are those remaining at the network interface after the use of the error control methods incorporated by the network.

TABLE VII
ASSESSMENT METHODS AND CONDITIONS FOR COMPRESSION TESTING

Sequence class	Reference format	Display (1)	Assessment method	Anchor condition
A	no reference	CIF	SS	H.263 at the same bitrate
B	no reference	CIF	SS	H.263 at the same bitrate
C	ITU-R 601	60 Hz ITU-R 601	DSCQS	MPEG-1 at the same bitrate MPEG-1 at the same bitrate MPEG-1 at the same bitrate
D	-	-	Evaluation by experts	-
E	CIF	CIF	DSIS	H.263 at the same bitrate H.263 at the same bitrate MPEG-1 at the same bitrate

(1) CIF displayed in a window centered in the 60 Hz ITU-R format.

TABLE VIII
SOURCE SEQUENCES AND VIDEO CHANNEL BITRATES FOR ERROR ROBUSTNESS TESTS

Sequence class	Video test sequences	Video bitrate (kbps)
Class A	Mother & Daughter; Akiyo; Hall Monitor; Container Ship	24
Class B	Silent Voice; Foreman; News; Coast Guard	48
Class C	Fun Fair; Table Tennis; Stefan; Mobile & Calendar	512

last nine of the ten decoded seconds were displayed during the subjective tests. This was done to provide reliable and fair viewing conditions by avoiding the influence of start-up effects, notably for low bitrates and low initial delays.

a) Test sequences and bit rates: For each sequence class, four source sequences were selected for testing, as shown in Table VIII.

b) Encoder constraints: Proposers were given the suggestion to use the following guidelines for the encoder constraints [5]:

- initial video codec delay[3]: 1.0 s maximum;
- instantaneous video codec delay[4]: 500 ms maximum (excluding first picture);
- average video codec delay[5]: 250 ms maximum.

These delay constraints imply the following:

- maximum number of bits used per frame: video encoded bit rate/2 (excluding the first frame);
- instantaneous encoded frame rate[6]: 2 f/s minimum;
- average encoded frame rate[7]: 4 f/s minimum.

Other encoder constraints could, however, be used provided that they were specified in the proposal technical description.

c) Test restrictions: Restrictions on the type of error control methods that were subjectively tested were as follows.

1) The decoder could only use information from the encoded bitstream. No out-of-band information could be used.

[3] Initial video codec delay is the instantaneous video codec delay for the first frame.

[4] Instantaneous video codec delay is the time from when the current picture is acquired to when the last bit for that picture is available in the bitstream to the decoder, accounting for a specific transmission bit rate, and assuming instantaneous acquisition, processing, and transmission times.

[5] Average video codec delay is the instantaneous video codec delay averaged over the entire sequence, excluding the initial video codec delay.

[6] Instantaneous encoded frame rate is the inverse of the time from the previous encoded frame to the current frame being encoded.

[7] Average encoded frame rate is the total number of encoded frames divided by the sequence length (in seconds), excluding the initial video codec delay.

TABLE IX
ERROR-PRONE VIDEO CHANNEL CONDITIONS

Test case	Residual error conditions	Description	Error interval [begin, end (s)]
1	10^{-3} Random Bit Error Rate	High Random BER	[1.5, end]
2	3 bursts of errors 50% BER within burst Random Burst Length: 16 to 24 ms Random bursts separation: > 2 s	Multiple burst errors	[1.5, 8]
3	Combination of test cases 1 and 2	High random BER and multiple burst errors	[1.5, end]

2) The testing procedure did not allow for feedback from the video decoder to encoder. This therefore precluded the testing of an automatic repeat request (ARQ) protocol for error control within a proposed codec.

Proposals that did not behave in such a way were still welcome, but they were evaluated in the same way as tools (evaluation by a panel of experts).

d) Error-resilience test conditions: The tested error conditions represent general characteristics of the residual errors that remain after the network has performed its error control. Each of the submitted compressed bitstreams was tested on all three residual error conditions defined in Table IX. Proponents were not allowed to make adjustments to their encoded bitstreams or decoder simulation after subjecting them to the exact error patterns. The exact error patterns were produced using the software in Annex E of the "MPEG-4 Testing and Evaluation Procedures Document" [5].

Each test case provided an initial period during which no errors were injected into the compressed bitstream, which allowed for the codec to transmit an initial frame and for the codec operation to stabilize into a steady state before errors were introduced. The time interval over which errors were introduced into a compressed bitstream is also shown in Table IX.

The three bursts in test cases 2 and 3 had lengths randomly selected from the range 16 to 24 ms. Their locations were also randomized, but they were separated by at least 2 s so that the effect of each burst could be observed.

e) Error-recovery test condition: This test condition was designed to judge the recovery performance of a codec following a very long burst error with loss of synchronization. This is needed for applications, such as broadcasting, in which a bitstream is provided without the possibility of a request to the encoder.

The test condition was a very long burst error followed by an error-free condition. The bit-error-rate during the burst error was 50%. The length of the burst error was randomly selected in the range of 1 to 2 s. The start of the burst was selected from the range 1.5 to 3 s. Proponents were not allowed to make adjustments to their encoded bitstreams or decoder simulation after subjecting them to the exact error patterns.

2) Test Methods: Formal subjective tests were used to evaluate both the error-resilience and the error-recovery performance of the codecs. The methods for each category are described below.

a) Error resilience test: The error-resilience of each codec was tested using the methods and conditions shown in Table X.

b) Error recovery test: The error-recovery test had the objective to evaluate the recovering capabilities of a codec in the presence of a long burst error. The error-recovery capability of each codec was tested in the same conditions as for the error-resilience tests in terms of bit rates and formats (see Table X).

The error-recovery capability was tested through a DSBV formal subjective test. The assessors were required to observe the decoded sequences without and with the effects of the burst errors (at the same bit rate) and they were then asked if the video quality in the error-prone sequence had recovered or not from the long burst error by the end of the sequence. Assessors were requested to answer "Yes" or "No" for each test sequence and for each bit rate.

V. PROCESSING OF RESULTS

In order to know about the relative merits of the video proposals presented to the MPEG-4 first round of tests, we will present in the next section the main results of the tests for each of the representative functionalities. In this section, the methodology used to process the results of the MPEG-4 tests is briefly described. The complete results are included in [6].

Since the scales used for the tests were categorical (items of quality or impairment were used), they had to be converted to numerical values as follows.

TABLE X
TEST METHODS AND CONDITIONS FOR ERROR ROBUSTNESS TESTS

Class of sequence	Bitrates	Library	Reference format	Display format (1)	Test method (error resilience)
A	24	ITU-R 601	no reference	CIF	SS
B	48	ITU-R 601	no reference	CIF	SS
C	512	ITU-R 601	ITU-R 601	60 Hz ITU-R 601	DSCQS

(1) CIF displayed in a window centered in the 60 Hz ITU-R 601 format.

- *Single Stimulus using a 0–10 score*: discrete score from 0 (bad) to 10 (excellent).
- *Double Stimulus using a five-note impairment scale*: discrete score from 1 (very annoying) to 5 (imperceptible).
- *Double Stimulus using a continuous quality scale*: for each case of the pair, scores were converted to a discrete value ranging from 0 (bad) to 100 (excellent). They were read for case A and case B separately, as a first step of processing. For each pair of votes, the difference was taken: reference sequence vote minus processed sequence vote. After this preprocessing, the best algorithms have the lowest scores; a number close to zero means that there is no difference between the reference and the assessed algorithm.
- *Double Stimulus using a binary vote*: the answers "YES" were converted to 1, and the answers "NO" to 0.

After, whatever the test method, and for each combination of algorithm and sequence, the average over observers of these preprocessed scores was calculated as well as the standard deviation and the confidence interval. Finally, for each case, the average over observers and sequences was calculated, together with the appropriate standard deviation and confidence interval given at 0.05 of minimum acceptable error on the mean.

The results for each bit rate were presented in tables, with the rows denoting the algorithms and anchors and the columns denoting the sequences. In an extra column, the average over sequences was given for each algorithm and anchor. The cells of the table were filled with the appropriate averages, standard deviations, and confidence intervals [6].

During the statistical processing of the data, a rejection criterion of inconsistent observers was applied. This filter was compliant with Recommendation ITU-R BT 500-6. However, no observer was rejected at the MPEG-4 tests using this filter. This indicates the particularly stable behavior of the observers throughout the sessions and within each session.

A calculation of an indicator of the distribution law (Beta 2 test) was also performed. It has shown that the global distribution of the scores was always approximately Gaussian though, in the DSCQS tests, the score range used by the test subjects was not very large. This limited range was due to the substantial difference in subjective quality between the reference and the coded sequences. The centered distribution of the data indicates that at least the parts of the rating scales used in the tests were well adapted to the tests [13]. Only the DSBV tests seemed to be not selective enough. These tests exhibited both a very high overall mean opinion score and a high overall standard deviation. This is easily explained by the scoring method, which constrained the scores to a binary choice (Y/N). This scale is definitely insufficient to provide reliable rankings for these tests, making it difficult to find statistically significant differences among the proposals.

An analysis of variance was performed systematically on each test session. The analyzed factors were the sequence effect, the proposal effect, and the observer effect. This was carried out to check for unwanted sequence dependent behavior and observer dependent behavior. This analysis verified that it is consistently the proposal effect that mainly explains the overall variance in the mean opinion scores. The sequence and observer variance within a test for sessions of the same size was reasonably stable.

When the grand means—global means for each session—are compared, no particular trend is detected. This fact leads to two conclusions: first, there were no fatigue and no learning effects evident in the test data, even for the tests that consisted of six separate test sessions; moreover, the random ordering of the proposals and sequences used to build each test session was effective.

There was very good coherence among the test sessions in terms of spread of video quality, which is a basic condition of no contextual effect in the test data. This is especially important in the case of the single stimulus test method. It was also observed that the test subject variance was always significant, but at a lower level than the other studied factors. This is very often the case in video subjective assessments, but in this test data it seems higher than usual. This is certainly due to the use of expert observers; such observers' perceptions of quality are typically found to be noticeably different from one subject to another, due to pollution by recognition of the basic techniques used to code, though one subject's judgement is stable from session to session.

Concerning the test data for different sequences, a Student comparative analysis was performed on the sequence results

TABLE XI
LIST OF ALGORITHMS PROPOSED TO THE NOVEMBER 1995 MPEG-4 TESTS

	Proposer - responsible person, company	MPEG doc. with technical description	Functionality addressed by the proposal
1.	Anastassiou, Columbia Univ.	MPEG95/374	Comp. A
2.	Brailean, Motorola	MPEG95/324	Comp. A, B Err. Res. A, B; Err. Rec. A, B
3.	Haskell, AT&T	MPEG95/365 MPEG95/486 MPEG95/442 MPEG95/423	Comp. A, B Comp. C Obj. Scal. A (alg.1), Obj. Scal. A (alg.2), B/C
4.	Lee, Microsoft	MPEG95/467	Comp. B, C, E Obj. Scal. A, B/C, E Spat. Scal. B/C/E
5.	Neff, UC Berkeley	MPEG95/317	Comp. A
6.	Talluri, TI	MPEG95/418	Comp. A, B Obj. Scal. A Err. Res. A, B; Err. Rec. A, B
7.	Nakaya, Hitachi	MPEG95/312	Comp. A, B
8.	Shimizu, JVC	MPEG95/315	Comp. A
9.	Etoh, Matsushita	MPEG95/393	Comp. B, E Obj. Scal. E Spat. Scal. B/C/E
10.	Machida, Matsushita	MPEG95/422	Err. Res. A, B Err. Rec. A, B
11.	Ito, Mitsubishi	MPEG95/341	Comp. B
12.	Asai, Mitsubishi	MPEG95/340	Comp. A Obj. Scal. A
13.	Miyamoto, NEC	MPEG95/338	Comp. A, B, Obj. Scal. A
14.	Jozawa, NTT	MPEG95/421	Comp. B, C
15.	Miki, NTT DoCoMo	MPEG95/403	Err. Res. A, B; Err. Rec. A, B
16.	Nakai, OKI	MPEG95/330	Err. Res. A, B
17.	Kuroe, OKI	MPEG95/339	Comp. A, B, C, E
18.	Hibi, Sharp	MPEG95/375 MPEG95/376	Comp. A Err. Res. A
19.	Kusao, Sharp	MPEG95/377	Temp. Scal. A, B/C/E
20.	Ogata, Sony	MPEG95/347	Comp. C
21.	Watanabe, Toshiba	MPEG95/354 MPEG95/352 MPEG95/353	Obj. Scal. A, Spat. Scal. A Err. Res. A, B; Err. Rec. A, B
22.	Park, Daewoo	MPEG95/348	Comp. A
23.	Moon, Hyundai	MPEG95/462	Comp. A
24.	Seo, Samsung	MPEG95/396	Comp. A, B
25.	Ohm, T. Univ. Berlin	MPEG95/333	Comp. C
26.	De Lamcillieure, HHI	MPEG95/415	Obj. Scal. A, B/C, Spat. Scal. A, B/C/E
27.	Gerken, Univ. Hannover	MPEG95/359	Comp. A Obj. Scal. A
28.	Nitsche, Bosch	MPEG95/454	Err. Res. A, B; Err. Rec. A, B
29.	Vial, Thomson	MPEG95/504	Comp. C
30.	Corset, Philips	MPEG95/408	Comp. B, E
31.	Ebrahimi, EPFL	MPEG95/320	Comp. A, B Obj. Scal. A
32.	Fryer, Univ. Strathclyde	MPEG95/429	Comp. A, Err. Res. A; Err. Rec. A

[MPEG95/XXX] <author>, Doc. ISO/IEC JTC1/SC29/WG11 MPEG95/XXX, Dallas, November 1995

within each test (applied on means and standard deviations over the proposals for each sequence). This analysis showed that, for class A sequences, "Akiyo" was always the most discriminating sequence, and "Mother & Daughter" was the least discriminating sequence. For class B sequences, "News" was always the most discriminating sequence, and for class C it was systematically "Stefan." Concerning the criticality of the sequences, no sequence was found to be significantly more critical than another. The discriminating power was estimated from the Student analysis of the standard deviation of each sequence over the different proposals. When a sequence is more discriminating, the difference of scores between proposals is increased, so its standard deviation is also increased. The criticality effect was analyzed on the basis of the comparison of the overall scores of each sequence over the proposals (by Student analysis again). When a sequence is always more

TABLE XII
MAIN RESULTS FOR THE COMPRESSION TESTS

COMPRESSION							
Class A		*Class B*		*Class C*		*Class E*	
Bitrate Anchor (rank)	6 Best Proposals (rank)	Bitrate Anchor (rank)	6 Best Proposals (rank)	Bitrate Anchor (rank)	3 Best Proposals (rank)	Bitrate Anchor (rank)	3 Best Proposals (rank)
10 kbps Anchor H263 (2)	Mitsubishi (1)	24 kbps Anchor H263 (2)	Motorola (1)	320 kbps Anchor MPEG1 (2)	NTT (1)	48 kbps Anchor H263 (2)	Matsushita (1)
	AT&T (2)		Matsushita (1)		AT&T (2)		Microsoft (1)
	UC Berkelev (2)		NTT (2)		Sonv (3)		Philips (3)
	Motorola (2)		AT&T (2)	512 kbps Anchor MPEG1 (2)	NTT (1)	112 kbps Anchor H263 (1)	Matsushita (1)
	EPFL (2)		Samsung (3)		AT&T (2)		Microsoft (2)
	U.Hannover (2)		TI (3)		Thomson (2)		Philips (2)
24 kbps Anchor H263 (1)	Mitsubishi (1)	48 kbps Anchor H263 (1)	EPFL (1)	1024 kbps Anchor MPEG1 (1)	NTT (1)	320 kbps Anchor MPEG1 (2)	Matsushita (1)
	AT&T (2)		AT&T (1)		AT&T (2)		Philips (2)
	Motorola (2)		Motorola (1)		Sonv (2)		Microsoft (2)
	UC Berkelev (2)		Matsushita (2)				
	TI (2)		Philips (2)				
	EPFL (3)		TI (2)				
48 khns Anchor H263 (1)	Mitsubishi (1)	112 khns Anchor H263 (1)	Matsushita (1)				
	AT&T (1)		TI (1)				
	TI (2)		AT&T (1)				
	UC Berkelev (2)		NTT (2)				
	Motorola (3)		Motorola (2)				
	U. Hannover (3)		Samsung (2)				

TABLE XIII
MAIN RESULTS FOR THE OBJECT SCALABILITY TESTS

OBJECT SCALABILITY					
Class A		*Class B/C*		*Class E*	
Bitrate No Anchor	3 Best Proposals (rank)	Bitrate Anchor (rank)	3 Best Proposals (rank)	Bitrate Anchor (rank)	Only Two Proposals (rank)
48 khns	Mitsubishi (1)	1024 kbps Anchor MPEG1 (1)	FHI (1)	320 khns Anchor MPEG1 (2)	Matsushita (1)
	Toshiba (2)		AT&T (2)		Microsoft (2)
	AT&T (2)		Microsoft (3)		

critical than others, whatever the proposals, its average score is different from the others. If this effect is statistically significant it can be said that this particular sequence is globally the most critical.

VI. PRESENTATION OF RESULTS

The list of algorithm proposals submitted to the November 1995 MPEG-4 video subjective tests is presented in Table XI [6].[8] The third column indicates the MPEG document where the algorithm technical description may be found. Notice that most of the algorithms that performed well in the MPEG-4 first round of tests have their technical description included as a paper in this special issue.

The list of tools submitted for evaluation in November 1995 is included in [14]. A few additional algorithm proposals were submitted to the January evaluation phase but they were not submitted to formal subjective testing [15].

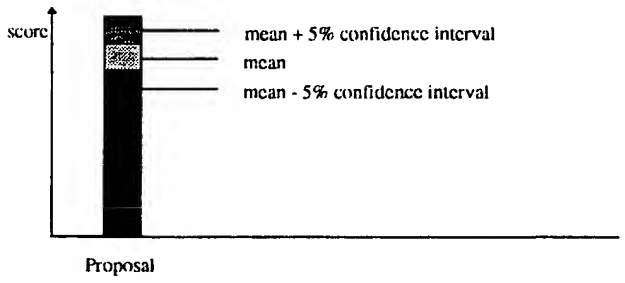
[8]Notice that no proposals were made for class C sequences in error robustness tests.

TABLE XIV
MAIN RESULTS FOR THE SPATIAL AND TEMPORAL SCALABILITY TESTS

SPATIAL SCALABILITY				TEMPORAL SCALABILITY			
Class A		*Class B/C/E*		*Class A*		*Class B/C/E*	
Bitrate No Anchor	Only Two Proposals (rank)	Bitrate No Anchor	3 Best Proposals (rank)	Bitrate No Anchor	Only One Proposal	Bitrate No Anchor	Only One Proposal
24 kbps	Toshiba (1)	512 kbps	Matsushita (1)	24 kbps	Sharp	512 kbps	Sharp
	HHI (2)		Microsoft (1)	24+24 kbps	Sharp	512+512kbps	Sharp
24+24 kbps	Toshiba (1)		HHI (2)				
	HHI (2)	512+512kbps	Matsushita (1)				
			HHI (1)				
			Microsoft (2)				

Excerpts of the experimental data from the tests are presented in a series of tables. There are table(s) for each of the tests listed in Section IV, but only approximately 40% of the overall results are provided, because when there is a large number of proposals in a test, only the first proposals are ranked. Tables XII–XVI give information on the relative performance of proposals, also compared to anchors when applicable. For each class of sequences, the first column in these tables gives the bit rate, the anchor used, and its relative ranking (in brackets), while the second column gives the proposer's company name and the proposal's relative ranking (in brackets). The ranking calculation was based on a complete statistical Student analysis, which provides a probability of equivalence for each pair of proposals [16]. The criteria used to decide that two proposals were statistically significantly different (SSD) was a probability of equivalence lower than 0.05. For example, if in the tables two proposals have the same ranking, this means that they were not statistically significantly different.

In addition to the tables described above, graphs of the average mean opinion score are provided (see Figs. 5–25). On each graph, the horizontal axis lists the proposals, and the vertical axis gives the mean opinion score (depending on the test method). The proposals are ordered by the mean opinion score and the proposal with the highest mean opinion score is at the far left of the graph. For every proposal displayed, a three-color bar is given including the confidence interval calculated for a probability of 5% as explained in the following graph.



VII. FINAL COMMENTS

It was not the purpose of this paper to comment on the technical merits or on the ranking in the tests of each proposal. However, from a subjective testing point of view, the first MPEG-4 tests provided very relevant indications to improve the specification of future tests.

During the first round of MPEG-4 video subjective tests, several test items (functionality, bit rate, class of sequence, etc.) were taken into account. Standard test methods were chosen and adapted according to the specific goal of each test. For example, at the highest bit rates, where the proposals were expected to have a quality level comparable to the original, the DSCQS method was used. In the case of the error recovery test, where the most important aspect seemed to be whether or not the algorithm under test recovered from the transmission errors, the DSBV method was used with a binary vote (instead of the standard DSIS assessment). For the evaluation of content-based scalability, both the global quality and the quality of a specific object in the scene were evaluated, in two separate tests.

From the comments of the test subjects, and from the outcome of the tests, a few possible improvements of the test methods have already been identified. Concerning the DSCQS method, the difference in quality between the sequences to be evaluated and the reference sequences was possibly too high in some cases. In such cases, the reference sequence may have been ineffective. There are two possible solutions to this problem: the first is to use an impaired reference instead of the original CCIR 601; the kind of impairment(s) which should be used to produce useful references must be studied and validated. The second solution is to use the single stimulus method for all bit rates, including the higher bit rates.

Concerning the error recovery test, nearly all of the proposals recovered from the errors by the end of the sequences (it was enough to intracode the last frame). Moreover the *recovery strategy* has a significant impact on the subjective quality and this was not assessed and reflected in this test. This basically means that the error recovery test performed did not allow the full evaluation of the recovery performance of the algorithms.

TABLE XV
MAIN RESULTS FOR THE ERROR RESILIENCE TESTS

ERROR RESILIENCE					
Class A			*Class B*		
Bitrate	Error Type	3 Best Proposals (rank)	Bitrate	Error Type	3 Best Proposals (rank)
24 kbps	Random BER 10E-3	Toshiba (1)	48 kbps	Random BER 10E-3	Motorola (1)
		TI (2)			TI (1)
		Motorola (2)			Toshiba (2)
	3 bursts	Toshiba (1)		3 bursts	TI (1)
		TI (2)			Matsushita (1)
		Matsushita (3)			Toshiba (1)
	Random BER 10E-3 + 3 bursts	Toshiba (1)		Random BER 10E-3 + 3 bursts	TI (1)
		TI (2)			Toshiba (2)
		Motorola (2)			Matsushita (2)

TABLE XVI
MAIN RESULTS FOR THE ERROR RECOVERY TESTS

ERROR RECOVERY			
Class A		*Class B*	
Bitrate	3 Best Prop.	Bitrate	3 Best Prop.
24 kbps	Matsushita (1)	48 kbps	TI (1)
	TI (1)		Matsushita (1)
	Toshiba (1)		Toshiba (1)

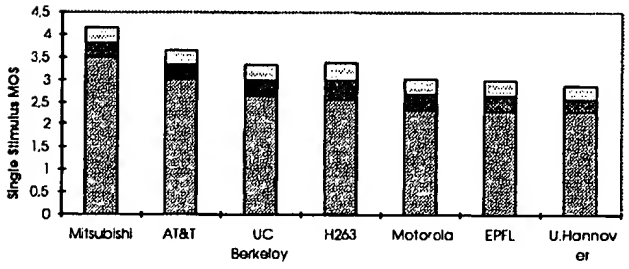


Fig. 5. Compression, class A, 10 kb/s.

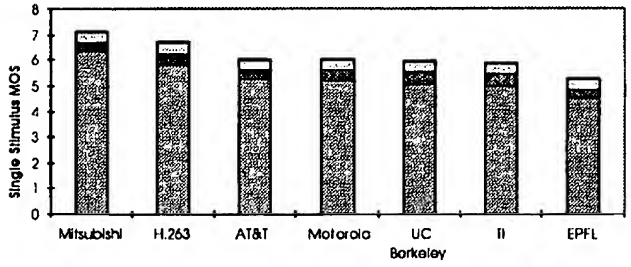


Fig. 6. Compression, class A, 24 kb/s.

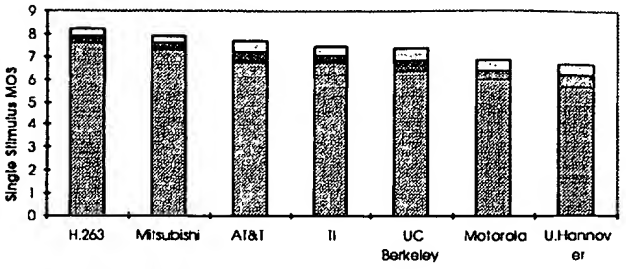


Fig. 7. Compression, class A, 48 kb/s.

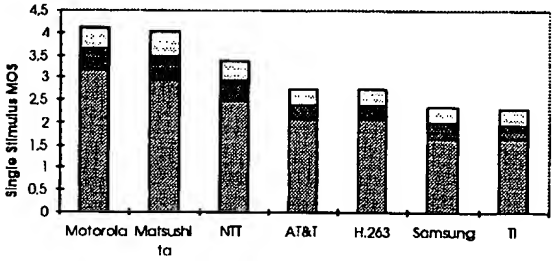


Fig. 8. Compression, class B, 24 kb/s.

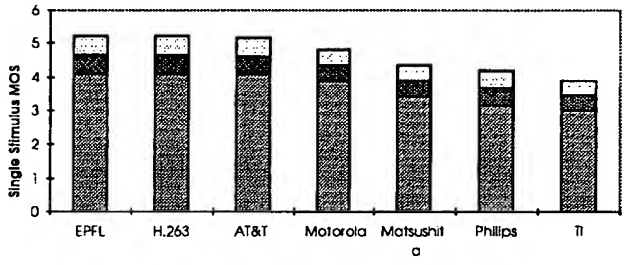


Fig. 9. Compression, class B, 48 kb/s.

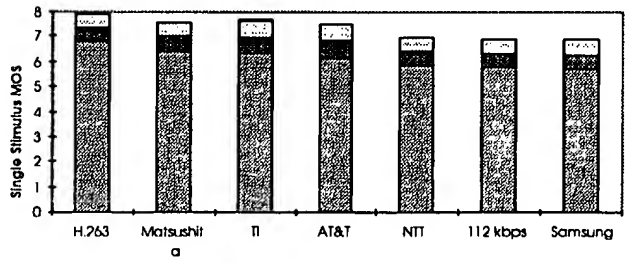


Fig. 10. Compression, class B, 112 kb/s.

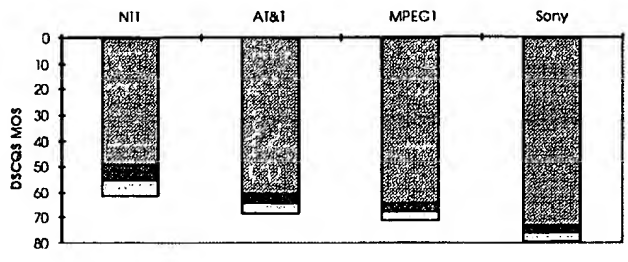


Fig. 11. Compression, class C, 320 kb/s.

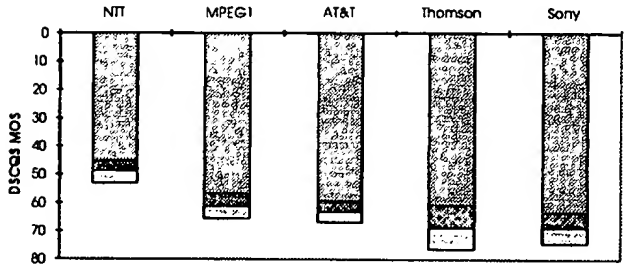


Fig. 12. Compression, class C, 512 kb/s.

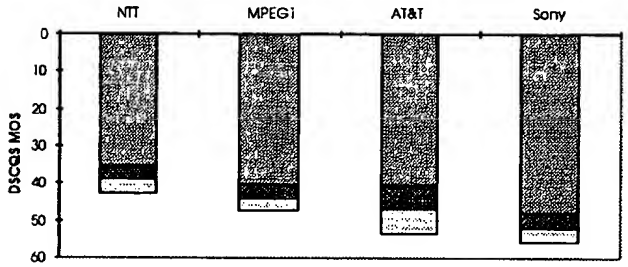


Fig. 13. Compression, class C, 1024 kb/s.

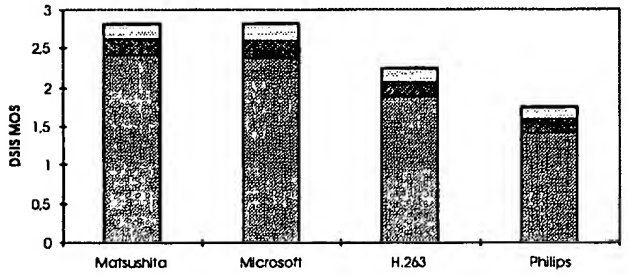


Fig. 14. Compression, class E, 48 kb/s.

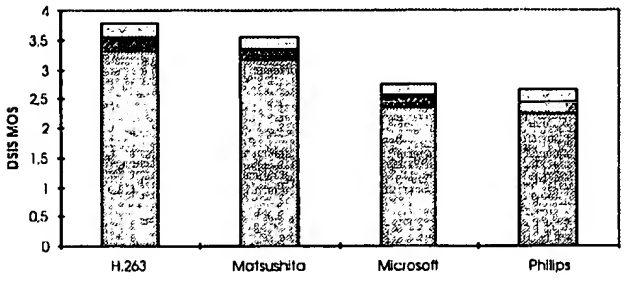


Fig. 15. Compression, class E, 112 kb/s.

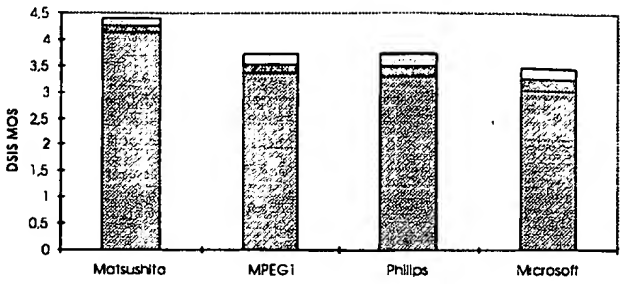


Fig. 16. Compression, class E, 320 kb/s.

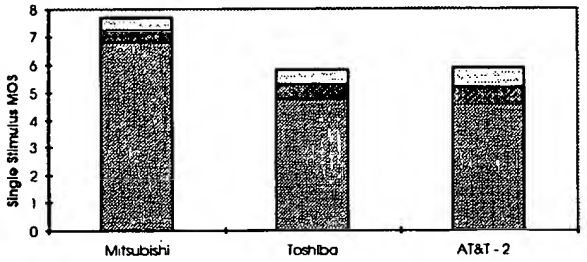


Fig. 17. Object scalability, class A, 48 kb/s.

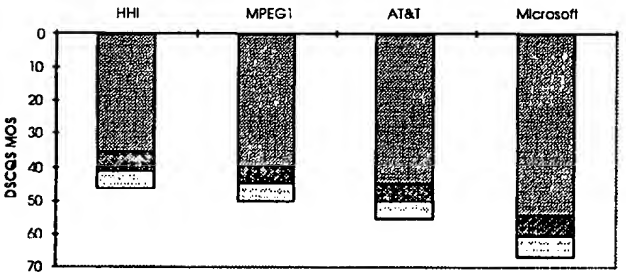


Fig. 18. Object scalability, class B/C, 1024 kb/s.

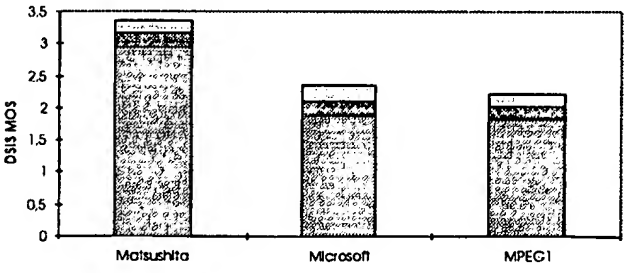


Fig. 19. Object scalability, class E, 320 kb/s.

An evaluation of the overall quality would have given more information; maybe a continuous evaluation using a slider could be a solution. The new method recently included in the ITU-R Rec. 500-7 could certainly overcome this problem; this method is the single stimulus using a continuous quality evaluation (SSCQE).

Finally, concerning content-based quality scalability, it is not clear that both global and object evaluations are really useful. In the case of global quality evaluation, the evaluation criteria can change from subject to subject, depending on how a subject's attention is divided between the background and the

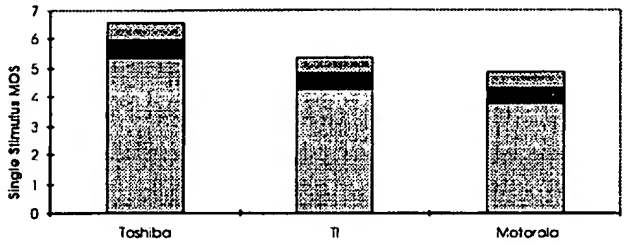


Fig. 20. Error resilience, 10E-3 random bit error rate, class A, 24 kb/s.

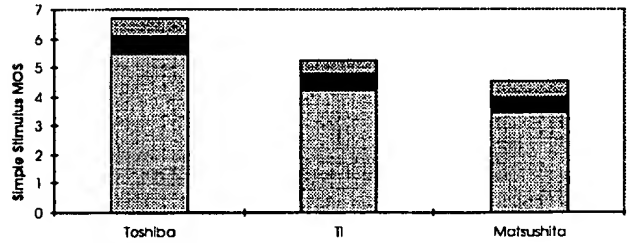


Fig. 21. Error resilience, three bursts of error, class A, 24 kb/s.

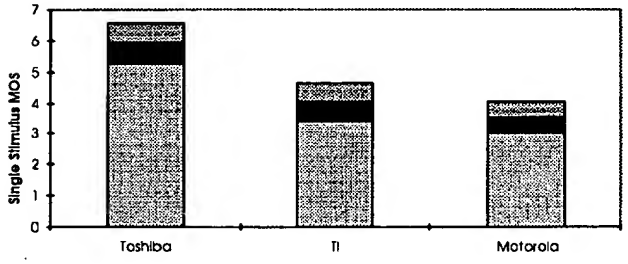


Fig. 22. Error resilience, three bursts of error + 10E-3 random bit error rate, class A, 24 kb/s.

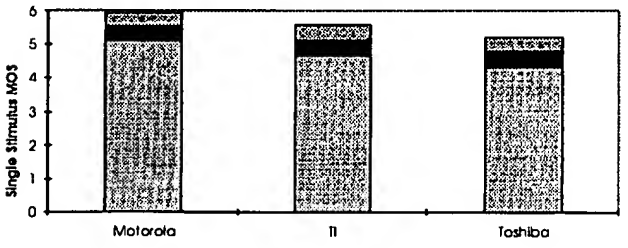


Fig. 23. Error resilience, 10E-3 random bit error rate, class B, 48 kb/s.

object. In the case of the object quality assessment, it may be more useful to extract the object from the sequence and show it on a uniform grey background. In fact, it seems the observers may be influenced, or at least distracted, by the background content and quality even when only the object quality has to be evaluated.

The MPEG-4 first round of video subjective tests was aimed at the comparison of several proposals and anchors in terms

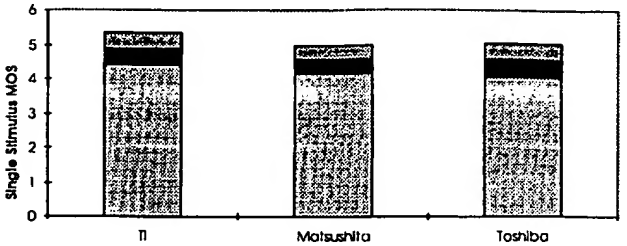


Fig. 24. Error resilience, three bursts of error, class B, 48 kb/s.

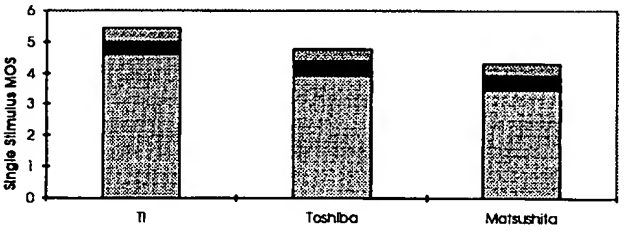


Fig. 25. Error resilience, three bursts of error + 10E-3 random bit error rate, class B, 48 kb/s.

of subjective quality. This objective was achieved since the outcome of the tests provided the MPEG video experts with the relevant information they needed to define the first MPEG-4 video verification model [7].

ACKNOWLEDGMENT

The authors would like to acknowledge all the MPEG-4 members for the interesting and fruitful exchange of opinions in the meetings and in the e-mail reflectors which very much enriched the authors' technical knowledge. Special thanks go for H. Peterson, L. Contin, V. Baroncini, J.-P. Thomas, and R. Koenen.

REFERENCES

[1] MPEG AOE Group, "Proposal package description (PPD)—revision 3," Doc. ISO/IEC JTC1/SC29/WG11 N998, Tokyo, July 1995.
[2] L. Chiariglione, "MPEG-4 project description," Doc. ISO/IEC JTC1/SC29/WG11 N1177, Jan. 1996.
[3] F. Pereira, "MPEG-4: a new challenge for the representation of audio-visual information," presented at *Picture Coding Symp. '96*, Melbourne, Australia, Mar. 1996.
[4] MPEG Systems Group, "Systems Working Draft, version 2.0," Doc. ISO/IEC JTC1/SC29/WG11 N1483, Maceió, Nov. 1996.
[5] F. Pereira, Ed., "MPEG-4 testing and evaluation procedures document," Doc. ISO/IEC JTC1/SC29/WG11 N999, Tokyo, July 1995.
[6] H. Peterson, "Report of the ad hoc group on MPEG-4 video testing logistics," Doc. ISO/IEC JTC1/SC29/WG11 MPEG95/532, Dallas, Nov. 1995.
[7] MPEG Video Group, "MPEG-4 video verification model 5.0," Doc. ISO/IEC JTC1/SC29/WG11 N1469 Maceió, Nov. 1996.
[8] Recommendation ITU-R BT 812, "Subjective assessment of the quality of alphanumeric and graphic pictures in teletext and similar services," 1994.
[9] Recommendation ITU-R BT 500-6, "Method for the subjective assessment of the quality of television pictures," 1994.
[10] EBU Report on Recovery Time, GT V1 2651, 1994.
[11] G. Bjontegaard *et al.*, "H.263 anchors—technical description," Doc. ISO/IEC JTC1/SC29/WG11 MPEG95/322, Dallas, Nov. 1995.

[12] ISO/CEI DIS 11172-2 (MPEG-1), "Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s."
[13] W. S. Togerson, *Theory and Methods of Scaling*. New York: Wiley, 1958.
[14] J. Ostermann, "Report on the ad hoc group on the evaluation of tools for non tested functionalities of video submissions," Doc. ISO/IEC JTC1/SC29/WG11 N1064, Dallas, Nov. 1995.
[15] ——, "Report on the ad hoc group on the evaluation of tools for non tested functionalities of video submissions for MPEG-4 in Jan. 1996," Doc. ISO/IEC JTC1/SC29/WG11 N1162, Munich, Jan. 1996.
[16] W. E. Duckworth, *Méthodes Statistiques de la Recherche Technologique*, Dunod, 1973.

Thierry Alpert was born in Paris, France, in October 1959. He graduated in biological and medical engineering and in signal processing at the University of Paris.

He joined the Centre Commun d'Etudes de Télédiffusion et Télécommunications in 1988 where he is currently in charge of the Image Quality Laboratory. He is also actively participating in several European projects and Standardization Committees (ACTS, ITU-R, ISO/MPEG, EBU, . . .). His work is focused on objective and subjective visual quality aspects mainly related to services based on digital image communication.

Fernando Pereira (S'88–M'90), for a photograph and biography, see this issue, p. 4.